UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS None
2a. SECURITY CLASSIFICATION AUTHORITY 2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution is unlimited.
4. PERFORMING ORGANIZATION REPORT NUMBER(S) NORDA Report 147	5. MONITORING ORGANIZATION REPORT NUMBER(S) NORDA Report 147
6. NAME OF PERFORMING ORGANIZATION Naval Ocean Research and Development Activity	7a. NAME OF MONITORING ORGANIZATION Naval Ocean Research and Development Activity
6c. ADDRESS *(City, State, and ZIP Code)* Ocean Science Directorate NSTL, Mississippi 39529–5004	7b. ADDRESS *(City, State, and ZIP Code)* Ocean Science Directorate NSTL, Mississippi 39529–5004
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Defense Mapping Agency/RE 8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS *(City, State, and ZIP Code)* Washington, DC 20305	10. SOURCE OF FUNDING NOS. PROGRAM ELEMENT NO.: 64701B PROJECT NO.: TASK NO.: WORK UNIT NO.:

11. TITLE *(Include Security Classification)*
Collecting Requirements for an International Geographic Names Processing System

12. PERSONAL AUTHOR(S)
Gail Langran

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT *(Yr., Mo., Day)*	15. PAGE COUNT
Final	From ____ To ____	October 1986	44

16. SUPPLEMENTARY NOTATION

17. COSATI CODES		
FIELD	GROUP	SUB. GR.

18. SUBJECT TERMS *(Continue on reverese if necessary and identify by block number)*
geographic names, data base, word processing

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

This document describes the requirements, future uses, interfaces, and data characteristics of a digital geographic names processing system (GNPS) for the Defense Mapping Agency (DMA). The GNPS will be comprised of a digital data base management system for names and their toponymic attributes, and a foreign text processor that will allow transliterated text to be typed, edited, and displayed in hard- and softcopy.

The work described here provided the basis for a Request for Proposal by the Defense Mapping Agency. Section 1 describes DMA's current names processing environment, its shortcomings, and past efforts to bring digital technology to DMA's names analysts. Section 2 defines the desired capabilities of the system to be procured. Section 3 discusses possible technical solutions to some GNPS requirements and the interactions of NORDA with the GNPS's future users. Section 4 summarizes the effort and describes the current project status. Section 5 cites references.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ■ DTIC USERS ☐	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22a. NAME OF RESPONSIBLE INDIVIDUAL Charles Walker	22b. TELEPHONE NUMBER *(Include Area Code)* (601) 688-4608 22c. OFFICE SYMBOL Code 351

DD FORM 1473, 83 APR EDITION OF 1 JAN 73 IS OBSOLETE.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

Naval Ocean Research and Development Activity
October 1986 NORDA - **Report 147**

Collecting Requirements for an International Geographic Names Processing System

Sponsored by Defense Mapping Agency/RE

Gail Langran
Mapping, Charting, and Geodesy Division
Ocean Science Directorate

Foreword

NORDA's Mapping, Charting, and Geodesy Division provides technical support to the Defense Mapping Agency in solving problems concerning automation of mapping and charting practices. Among the problems that have resisted automation is the processing of geographic names. This study collected requirements for an off-the-shelf digital system to enter, store, retrieve, sort, and format geographic names with diacritics.

A. C. Esau

A. C. Esau, Captain, USN
Commanding Officer, NORDA

Executive summary

This document describes the requirements, future uses, interfaces, and data characteristics of a digital geographic names processing system (GNPS) for the Defense Mapping Agency (DMA). The GNPS will be comprised of a digital data base management system for names and their toponymic attributes, and a foreign text processor that will allow transliterated text to be typed, edited, and displayed in hard- and softcopy.

The work described here provided the basis for a Request for Proposal by the Defense Mapping Agency. Section 1 describes DMA's current names processing environment, its shortcomings, and past efforts to bring digital technology to DMA's names analysts. Section 2 defines the desired capabilities of the system to be procured. Section 3 discusses possible technical solutions to some GNPS requirements and the interactions of NORDA with the GNPS's future users. Section 4 summarizes the effort and describes the current project status. Section 5 cites references.

NOTE: A number of hardware and software companies are mentioned in this report to illustrate current off-the-shelf technical capabilities. Such references should not be construed as endorsement of these companies' products by either the author or the government.

Acknowledgments

This work proceeded quickly and smoothly because of the extremely effective coordination of Ferne Cooper, the DMAHTC/SDST project engineer. Ms. Cooper arranged and attended all GNPS working group meetings and responded immediately to all requests for additional information and materials. Answers were produced so quickly and reliably that I was encouraged to dig for greater detail, with accordingly improved results. The DMAHTC/SDSN working group, which patiently described their needs and supplied feedback on the work presented in this report, deserve a great deal of credit for their efforts. The working group was comprised of Fikry Galal, Charles Heyda, Gerd Quinting, Frederick Rohrer, and Edward Szymanski. Randall Flynn's involvement was also instrumental to the project. Mr. Flynn (of DMAHTC/SDT) read the compiled results of the NORDA study, provided many useful comments, and worked with Ms. Cooper to transform the NORDA study into a Request for Proposal acceptable to the DMAHTC contracts office. His review of and comments on this report were extremely helpful. This study was funded by DMAHQ/STT (Program Element 64701B); its project manager was Dennis Franklin.

Contents

Collecting requirements for an International Geographic Names Processing System

1. DMA's current names processing environment

INTRODUCTION

The Defense Mapping Agency (DMA) is responsible for collecting and maintaining a correct file of geographic placenames worldwide to support a variety of U.S. government activities. All placenames must be documented with an array of data that includes feature type, position, and possibly some feature attributes; all names previously or currently associated with the feature; reference sources used to acquire the names data; and information concerning the Board on Geographic Names' Foreign Names Committee ruling on the feature's names. Currently, all such data is recorded on approximately 4.5 million paper index cards via pen, pencil, or typewriter.

DMA plans to augment the card file with a digital geographic names data base. Working with placenames data in a digital environment, however, means that hardware and software must be available to enter, view, edit, and print the diacritics and special symbols used in foreign text. Lack of such hardware and software has been a major stumbling block in establishing an operational digital data base for DMA's toponymic use.

This report describes the needs, uses, interfaces, and data characteristics of a digital geographic names processing system (GNPS) for DMA. The GNPS will manage a names data base and provide software for foreign text processing. GNPS requirements were defined by a cooperative NORDA/DMA effort during the first quarter of FY 86.

The rest of this section describes DMA's current names processing environment. Section 2 summarizes the GNPS's operational requirements. Section 3 discusses some possible technical solutions to GNPS operational requirements and the methods used to derive them. Section 4 summarizes project progress. Section 5 lists the references cited in this report.

DMA'S NAMES PROCESSING RESPONSIBILITIES

DMA's names analysts are linguists and geographers, each of whom specializes in a region of the world. The names analysts supply names for DMA maps and charts, and confirm that the mapped depiction of political boundaries and sovereignties conform with U.S. position, thereby establishing an important liaison between the U.S. State Department and DMA. They also publish printed lists of feature names, called gazetteers (Table 1–1). Finally, DMA's toponymists respond to inquiries from other government agencies concerning geographic names and make recommendations regarding name changes to the Board on Geographic Names.

Thus, DMA's names analysts support production but also must perform a great deal of ongoing research. Names reflect U.S. recognition or nonrecognition of an area's political status, making them vitally important in sensitive areas. Toponymic research involves comparing many foreign-language sources, transliterating names in non-Roman alphabets, and Romanizing names from Oriental ideographies. Changes in political regimes can cause thousands of name changes: over 100,000 name changes have occurred in the USSR since 1920; 30,000 names have changed in Poland; and many African states have comparable rates of upheaval (DMAHTC, 1978).

SHORTCOMINGS OF TODAY'S ANALOG METHODS

As mentioned earlier, DMA's names processing activities revolve around a file of 4.5 million paper index cards called the Foreign Place Names File (FPNF). Approximately 3 million FPNF cards are feature cards: each contains the names data for a single feature, which includes a primary name and alternate names (if any exist) and their sources. Figure 1–1 is a facsimile of an FPNF feature card, front and back. The remaining 1.5 million cards are cross references to the feature cards. Most FPNF entries are handwritten; if entries are typed, diacritics are added by hand.

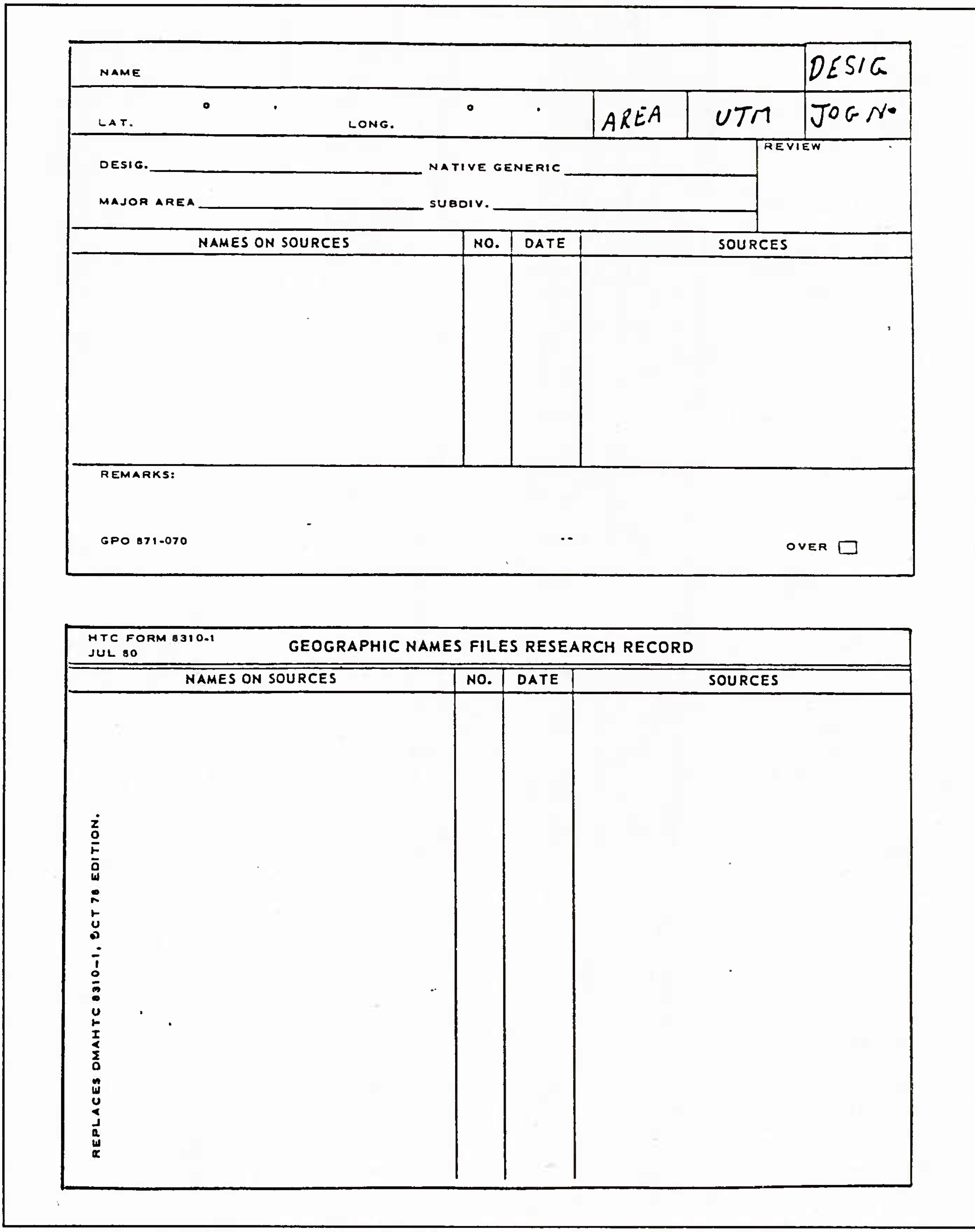
Minimal technical support has been provided to ease the task of DMA's names analysts. The manual methods employed today are slow, and thumbing through an alphabetized card file misuses a professional toponymist's time. Also, FPNF data is not easily correlated with map and chart names, meaning that no safeguards exist against a single feature mistakenly being given different names on different maps, charts, and gazetteers.

Table 1–1. Sample pages from a DMA gazetteer.

NAME		DESIG.	LAT.	LONG.	AREA	UTM	JOG NO.
Bivio Fiumbiro: see Funyan Bīra		PPL	9°24'N	42°19'E	ET07	KF04	NC38-09
Bīwot Mīka'ēl Bēte Kristīyan		CH	11°37'N	39°13'E	ET14	EC28	NC37-03
Bīws		PPL	12°36'N	39°06'E	ET14	ED19	ND37-15
Biyagundi	*	PPL	14°28'N	37°12'E	ET04	CF09	ND37-05
Biyanot: see Bīye Ānod		PPL	10°36'N	42°40'E	ET07	KG47	NC38-05
Bīyara		PPL	14°29'N	37°57'E	ET12	CG80	ND37-06
Bīye Ābī		PPL	10°24'N	34°43'E	ET13	XG85	NC36-08
Bīye'ādē YeTerara Ch'af		PK	9°20'N	42°34'E	ET07	KF33	NC38-09
Bīye Ānod		PPL	10°36'N	42°40'E	ET07	KG47	NC38-05
Bīye Bahī		PPL	10°03'N	42°30'E	ET07	KG21	NC38-05
Bīye Denan Shet'		STMI	10°02'N	42°23'E	ET07	KG11	NC38-05
Bīye Gurgur		PPL	10°24'N	42°41'E	ET07	KG45	NC38-05
Biye Gurgur Polīse T'abiya: *see* Bīye Gurgur Polīs T'abīya		PP	10°24'N	42°41'E	ET07	KG45	NC38-05
Bīye Gurgur Polīs T'abīya		PP	10°24'N	42°41'E	ET07	KG45	NC38-05
Bīye Gurgur Shet'		STMI	10°35'N	42°40'E	ET07	KG47	NC38-05
Bīyehun Shet'		STM	9°46'N	42°22'E	ET07	KF18	NC38-09
Biye K'obē: *see* Bīye K'obē Polīs T'abīya		PP	10°23'N	42°34'E	ET07	KG34	NC38-05
Bīye K'obē Polīs T'abīya		PP	10°23'N	42°34'E	ET07	KG34	NC38-05
Bīyo		PPL	8°28'N	40°33'E	ET07	FV73	NC37-16
Bīyo		PPL	8°36'N	40°01'E	ET01	FV15	NC37-15
Bīyo		PPL	9°17'N	42°00'E	ET07	JF72	NC37-12
Bīyo		PPL	9°18'N	42°01'E	ET07	JF72	NC38-09
Bīyo		PPL	9°28'N	38°22'E	ET10	DA34	NC37-10
Bīyo		PPL	9°28'N	42°42'E	ET07	KF44	NC38-09
Bīyo		PPL	10°46'N	39°00'E	ET14	EB09	NC37-06
Biyo (1): see Bīyo		PPL	9°17'N	42°00'E	ET07	JF72	NC37-12
Biyo (2): see Bīyo		PPL	9°18'N	42°01'E	ET07	JF72	NC38-09
Bīyo Karaba		PPL	9°22'N	41°10'E	ET07	GA33	NC37-12
Bīzara Bota		LCTY	11°34'N	39°34'E	ET14	EC67	NC37-03
Bīzeb		PPL	11°45'N	37°30'E	ET06	CC39	NC37-01
Bizen, Monte: see Bīzen Terara		MT	15°20'N	39°06'E	ET04	EG19	ND37-03
Bīzen Terara		MT	15°20'N	39°06'E	ET04	EG19	ND37-03
Bizet	*	PPL	14°23'N	39°15'E	ET12	EF28	ND37-07
Bīzhī Shet'		STM	7°27'N	35°02'E	ET08	YD22	NB36-04
Bīzīdīmo Mīka'ēl Bēte Kristīyan		CH	9°10'N	36°51'E	ET13	BA61	NC37-09
Blanga	†*	PPL	4°44'N	40°04'E	ET11	FR12	NB37-15
Blue Nile [conventional]; Ābay Wenz [ETHIOPIA]; Al Baḩr al Azraq [SUDAN]		STM	15°38'N	32°31'E	ET00	VN42	ND36-02
Blue Nile Falls: *see* T'īs Isat Fwafwatē		FLLS	11°29'N	37°35'E	ET02	CC46	NC37-02
Boa	†*	PPL	3°41'N	38°44'E	ET11	DQ70	NA37-02
Bo'a Ts'a'ida		PPL	14°23'N	38°34'E	ET12	DF59	ND37-06
Boba		PPL	8°07'N	36°17'E	ET08	BU09	NC37-13
Bobbe	*	MTS	6°54'N	37°18'E	ET09	CT16	NB37-05
Bobbe, Monti: see Bobbe		MTS	6°54'N	37°18'E	ET09	CT16	NB37-05
Bobē		PPL	9°00'N	38°29'E	ET10	DV49	NC37-14
Bobē		PPL	9°30'N	38°15'E	ET10	DA15	NC37-10

Table 1–1. Cont'd.

Boren		PPL	10°27'N	39°25'E	ET10	EB45	NC37-07
Boren		PPL	12°07'N	39°45'E	ET14	ED83	ND37-15
Boren		PPL	12°10'N	39°45'E	ET14	ED84	ND37-15
Boren		PPL	12°11'N	39°45'E	ET14	ED84	ND37-15
Boren (1): see Boren		PPL	12°11'N	39°45'E	ET14	ED84	ND37-15
Boren (2): see Boren		PPL	12°10'N	39°45'E	ET14	ED84	ND37-15
Boren (3): see Boren		PPL	12°07'N	39°45'E	ET14	ED83	ND37-15
Borena		PPL	10°45'N	38°46'E	ET14	DB78	NC37-06
Borena and Saynt: see Borena Āwraja		ADM2	10°50'N	38°45'E	ET14	DB79	NC37-06
Borena Āwraja		ADM2	4°40'N	40°00'E	ET11	FR11	NB37-15
Borena Āwraja		ADM2	10°50'N	38°45'E	ET14	DB79	NC37-06
Borena Bota		LCTY	4°57'N	38°19'E	ET11	DR24	NB37-14
Borena Bota		LCTY	5°22'N	39°27'E	ET11	ER49	NB37-11
Borena Bota		LCTY	10°42'N	38°43'E	ET14	DB68	NC37-06
Borenea: see Borena Āwraja		ADM2	4°40'N	40°00'E	ET11	FR11	NB37-15
Borē Shet'		STMI	8°46'N	38°28'E	ET10	DV46	NC37-14
Borey		PPL	9°30'N	38°59'E	ET10	DA95	NC37-10
Borgebba	*	PPL	7°57'N	33°35'E	ET08	WD67	NB36-03
Borgela		PPL	6°20'N	36°40'E	ET05	BT40	NB37-05
Borgianil	†*	PPL	6°14'N	43°39'E	ET07	LB58	NB38-06
Borgo: see Bargo		PPL	6°51'N	41°10'E	ET03	GT35	NB37-08
Borguddo	†*	PPL	4°20'N	39°08'E	ET11	EQ17	NB37-15
Borī		PPL	9°39'N	37°05'E	ET13	BA86	NC37-09
Bork'a Shet'		STM	7°27'N	40°57'E	ET03	GU12	NB37-04
Borkē: see Borkena Shet'		STM	10°36'N	40°25'E	ET10	FB57	NC37-07
Borkena Shet'		STM	10°06'N	39°59'E	ET10	FB01	NC37-07
Borkena Shet'		STM	10°36'N	40°25'E	ET10	FB57	NC37-07
Borkena Shet'		STM	10°52'N	40°04'E	ET14	FC10	NC37-07
Borkenna: see Borkena Shet'		STM	10°36'N	40°25'E	ET10	FB57	NC37-07
Borkoshē Bota		LCTY	6°50'N	37°55'E	ET11	CT85	NB37-06
Borle: see El Borle		WLL	5°04'N	43°31'E	ET03	LA36	NB38-10
Borle-ier	†*	PPL	5°06'N	42°26'E	ET03	KA16	NB38-09
Borlēy Shet'		STMI	9°52'N	41°42'E	ET07	GA99	NC37-12
Borni	†*	PPL	12°41'N	36°01'E	ET02	AE70	ND37-13
Boro		PPL	9°20'N	38°35'E	ET10	DA53	NC37-10
Boro		PPL	9°31'N	38°19'E	ET10	DA25	NC37-10
Boro		PPL	9°50'N	37°05'E	ET13	BA88	NC37-09
Boro	*	MT	11°22'N	39°39'E	ET14	EC75	NC37-03
Boro: see Boro Shet'		STM	9°21'N	38°36'E	ET10	DA53	NC37-10
Borò, Monte: see Boro		MT	11°22'N	39°39'E	ET14	EC75	NC37-03
Boro Āreda		PPL	10°01'N	38°46'E	ET10	DB70	NC37-06
Boro Bota		LCTY	8°49'N	38°32'E	ET10	DV47	NC37-14
Borodda: see Boreda		PPL	6°21'N	37°42'E	ET05	CT50	NB37-06
Borodda: see Boreda		PPL	6°32'N	37°46'E	ET11	CT62	NB37-06
Boroi: see Beroy		PPL	8°18'N	35°26'E	ET08	YE61	NC36-16
Boroli	*	MT	10°08'N	41°03'E	ET07	GB22	NC37-08
Boroli, Monte: see Boroli		MT	10°08'N	41°03'E	ET07	GB22	NC37-08


NAME | DESIG
LAT. ° ' LONG. ° ' | AREA | UTM | JOG No.

DESIG. ________ NATIVE GENERIC ________ | REVIEW

MAJOR AREA ________ SUBDIV. ________

NAMES ON SOURCES	NO.	DATE	SOURCES

REMARKS:

GPO 871-070

OVER ☐

HTC FORM 8310-1
JUL 80

GEOGRAPHIC NAMES FILES RESEARCH RECORD

NAMES ON SOURCES	NO.	DATE	SOURCES

REPLACES DMAHTC 8310-1, OCT 78 EDITION.

Figure 1-1. FPNF card facsimile.

The most serious shortcomings of DMA's current system, however, stem from DMA's impending overall modernization. Within the next five years most DMA production processes will be performed with the help of digital systems. Maps in progress will be reviewed and edited by cartographers using softcopy workstations and digital data. A number of problems could ensue if analog names processing continues. Once the speed of other mapping processes increases, names preparation will consume a proportionally large piece of the overall production time unless efficiency is improved. Additionally, if names data processing remains analog, plans must be made to integrate the names into the digital production pipeline.

SUMMARY OF EFFORTS TO INTRODUCE DIGITAL SYSTEMS

Evidently, an automated retrieval system to draw correct geographic names from data banks would be a far more efficient way to handle production, particularly in emergencies when quick response is critical. The complexity of the research required to maintain names makes designing such a retrieval system difficult. A great deal of diverse information must be stored and accessed in flexible ways, not merely compiled to standard formats. DMA also has ambitious goals for increasing the size of its names data base.

DMA has funded studies of the problem first at the U.S. Army Engineer Topographic Laboratories (ETL, 1983), then at NORDA (Brown et al., 1983; Langran et al., 1985a,b; Langran, 1985a–d). In FY 87 DMA plans to contract for the development of an operational digital geographic names processing system (GNPS) for delivery in FY 88. NORDA worked with DMAHTC in FY 86 to frame reasonable operational requirements given time, technology, and funding limitations, and developed prototype forms and menus that were acceptable to the GNPS's future users. The next section describes the GNPS's operational requirements. The methods used to deduce the requirements and some possible technical solutions are discussed in Section 3.

2. Future names processing requirements

THE NAMES DATA BASE

The FPNF, paper maps and charts, and gazetteers comprise DMA's analog names data base. Of these analog data sources, however, all are lacking one or more items of information vital to names processing. The FPNF and gazetteers have very poor positional accuracy: both truncate position to the nearest minute, providing a worst-case positional variance of over one mile near the equator (over one inch at 1:50,000 scale). Large-scale maps supply adequate positional accuracy but their names are notoriously difficult to capture digitally (Langran, 1985d), they are sometimes incorrect or out-of-date, and they lack the toponymic information (i.e., reference source, approval data) needed to maintain them. Evidently, populating the data base will be a major task entailing many years of work.

Estimates of the data base's ultimate size vary radically. The modal estimate, however, is 50 million names—the approximate number needed to describe all features shown on DMA's 1:250,000 topographic maps. Thus, the data base problem goes beyond the formidable task of converting a fully analog operation to digital. It also entails implementation of a data base management system (DBMS) whose data base will grow several orders of magnitude from the time of inception. Added to the bulk of the placenames is auxiliary data. The next three subsections discuss GNPS data needs.

Required data

Every complete data base record requires certain information to be stored concerning that name. A list of such entries follows.

Country. Names data would seem to partition itself naturally along international lines were it not for features that cross international boundaries (fewer than 4% of all features) and a continuing need to compile names for rectangular areas (maps) irrespective of international divisions. Evidently, each country has many features and each feature is located in a minimum of one country. But features crossing international boundaries are associated with more than one country, and may have a different name or series of names in each country. Today, 229 countries and dependencies have Federal Information Processing Standard (FIPS) codes. The longest country name is 38 characters, although most are considerably shorter.

Administrative area code. All populated places are assigned a standard 4-digit code that describes a feature's country and political subdivision (generally the state or province). The first two digits are the FIPS country code. Table 2–1 shows Ethiopia's area codes (generally called "ADM1s"). Populated places that cross administrative boundaries are assigned a special ADM1. The number of ADM1s per country ranges from one to 122. The system must be capable of displaying the ADM1 or the full subregion name (e.g., "Arsi" or "ET01").

Table 2–1. Area codes for Ethiopia (without diacritics).

ET00	ETHIOPIA	ET08	Ilubabor
ET01	Arsi	ET09	Kefa
ET02	Gonder	ET10	Shewa
ET03	Bale	ET11	Sidamo
ET04	Ertra	ET12	Tigray
ET05	Gamo Gofa	ET13	Welega
ET06	Gojam	ET14	Welo
ET07	Harerge		

Geographic position. Latitude and longitude in degrees, minutes, and seconds are needed for every named feature. Currently, all feature positions are recorded at one point only. An areal feature's position is stated as a point near the feature's center. Because most linear features are rivers and streams, their point positions are generally their mouths. UTM coordinates, needed for some applications, can be derived computationally from the latitude and longitude. A further positional descriptor used at DMA is the JOG sheet number, which is the number of the 1:250,000 topographic map upon which the feature would appear. JOG sheet number can also be derived via software from the latitude and longitude.

Feature designation. A feature's type (i.e., populated place, stream, etc.) must be stored. A standard feature designation code is used for gazetteers. A decoded feature designation should also be available for reference.

Placename. One or more proper names are stored for each geographical feature (e.g., one feature can have many names). Average placename lengths vary with countries, but these general rules apply: approximately 95% of all names are under 25 characters, 99% are under 40 characters. However, the GNPS must provide for a few names that exceed 120 characters; previous studies have suggested using an overflow field. A "wildcard character" capability is required of the GNPS DBMS so analysts can access names using partial spellings filled in with wildcards (e.g., find all populated placenames of any length that begin with "b," have a "g" in the middle, and end with "ham").

Type of name. Table 2–2 shows the different name types and their 1- to 3-digit codes (developed by DMA's toponymists). Both the complete and the coded name type should be accessible from the system. Short forms of names are generally nested within their long forms, but reference sources often differ. A small percentage of unapproved names are designated "not verified." These names are never included on gazetteers or maps, and must never be compiled unless expressly requested.

Reference source information. Every name must be accompanied by a source citation, to include title and date.

Table 2–2. Types of geographic names and their codes.

Type of Name	Code
Approved	A
Approved Local Official	L
Approved Local Official Short Form	LS
Approved Local Official Long Form	LL
Approved Conventional	C
Approved Conventional Short Form	CS
Approved Conventional Long Form	CL
Daggered Name	D
Pending Approval	A1
Pending Approval, Local Official	L1
Pending Approval, Local Official Short Form	LS1
Pending Approval, Local Official Long Form	LL1
Pending Approval, Conventional	C1
Pending Approval, Conventional Short Form	CS1
Pending Approval, Conventional Long Form	CL1
Variant	V
Variant, never approved	V1
Variant, formerly approved	V2
Not verified	NV

Optionally, a scale can be included for graphic sources or a page number for text sources. Names can have many source citations, although ten is a reasonable ceiling; one source can, of course, cite many names. When a name or a feature is omitted from a source, this information, sometimes important, must be stored.

Optional data

Different regions of the world and different types of named features cause differing demands in data storage among records. Estimates of the need for each data type have been made whenever possible.

Feature association. Analysts may wish to associate one feature with another. A simple reason is when two nearby features have the same name. If an analyst ascertains that two unique features exist, he logically associates them to prevent duplication of his research at a later date. Currently, fewer than 5% of all features are associated to another.

Feature research notes. The free format of the FPNF cards has allowed analysts to add notations to the bottom. Users agree that these notations can be extremely helpful and should be supported by the digital system. Short (e.g., 80-character) notes are acceptable, provided more than one note can be attached to each feature.

Approval of the names entry. All the names related to a single feature are referred to as an "entry." Complete names entries (not single names) are approved via a formal process that includes linguistic and geographic review, and action on the part of the Foreign Names

Committee (FNC). Thus, one feature can have more than one approved name. Every entry must cite its approval data: the FNC meeting number, the initials of the geographer and linguist, and the dates of the geographic and linguistic reviews. One set of approval data may apply to many names, since blanket approvals are often issued, particularly in the case of gazetteers. More than 80% of all names are part of an approved entry.

Form of name. In areas using non-Roman alphabets or ideographies it is important to know whether the name stored in the data base was transliterated by a DMA analyst or by the reference source originator.

Language of a name. In bilingual areas, a different name might be approved for each language.

Name research notes. Toponymists may wish to add free-form annotations to individual names. As with feature research notes, name research notes can be restricted to 80 characters each, provided that more than one note can be attached to each name.

Country of name. International feature names often differ from country to country. The country that uses a given name must be identified.

Conceptual model and data dictionary

Figure 2–1 shows the data base's conceptual model. Table 2–3 is a prototype data dictionary that defines the conceptual model's entity names.

Interfaces to the data base

The GNPS will be a stand-alone system. In the future, however, features stored in the GNPS data base must be matched to features stored in DMA's planned mapping, charting, and geodesy (MC&G) data base. Because a hardware link between the two data bases is not possible (and because the GNPS may have exceeded its life cycle before the MC&G data base is fully operational) a logical match is required.

Details of the MC&G data base structure are not widely available and, in some cases, are not firm. This study has devised two matching strategies; both may prove useful. The first is to provide space in the GNPS data base for a feature key. If the MC&G data base employs such a key, it can be inserted into the GNPS data base once a match is made between a GNPS feature and an MC&G feature, which would make future matches relatively simple. The initial match between the two features is, however, still a problem.

The second strategy involves matching other data elements in the feature records to progressively narrow the field of candidate matching features. Several data elements are stored in both data bases. The following evaluates their usefulness in making a software match between two features.

Matching countries. The country in which the feature is located can be used to narrow the search range, but features crossing international boundaries (e.g., located in several countries) will complicate the procedure.

Matching coordinate locations. Comparing the coordinate locations stored in the MC&G data base to those in the names data base will be difficult. At its inception, a high percentage of GNPS feature coordinates will be rounded to the nearest minute (cross-referencing between the two data bases would be an excellent way to acquire better coordinates for GNPS features). A second problem with coordinate matching will be matching the single coordinate points of linear and areal features stored in the names data base with the complete feature outlines stored in the MC&G data base.

Matching feature designations. The feature designation will be very useful to narrow the search range. Even though different feature codes are used by the two data bases, translation should not be difficult.

Matching feature names. The feature name that is used on DMA maps and charts will be stored in the MC&G data base. If a name has changed, it will probably be updated only in the GNPS data base unless procedures are developed to update the MC&G data base at the same time. In either case, the old form should remain in the GNPS data base for the software to find and reference to the feature.

Evidently, software cannot irrefutably match features between the two data bases without analyst support. A reasonable scenario would produce two output files: one with solid matches, the other with ambiguous matches that must be resolved interactively.

FOREIGN TEXT PROCESSING

Only a small subset of the world's languages use a Roman alphabet. Other alphabets include Arabic, Cyrillic, Hebrew, and Greek. The technical difficulties of digitally processing alternate alphabets are nominal. The technical difficulties of digitally processing the ideographic systems employed in East Asia, however, present a major challenge, since their ideographs number in the thousands. Efforts have been made in both Japan and China to reduce the number of ideographs in common use (to about 3500 and 7000, respectively) but placenames, particularly those derived from historical or remote sources, are likely to contain unusual characters not included in the streamlined

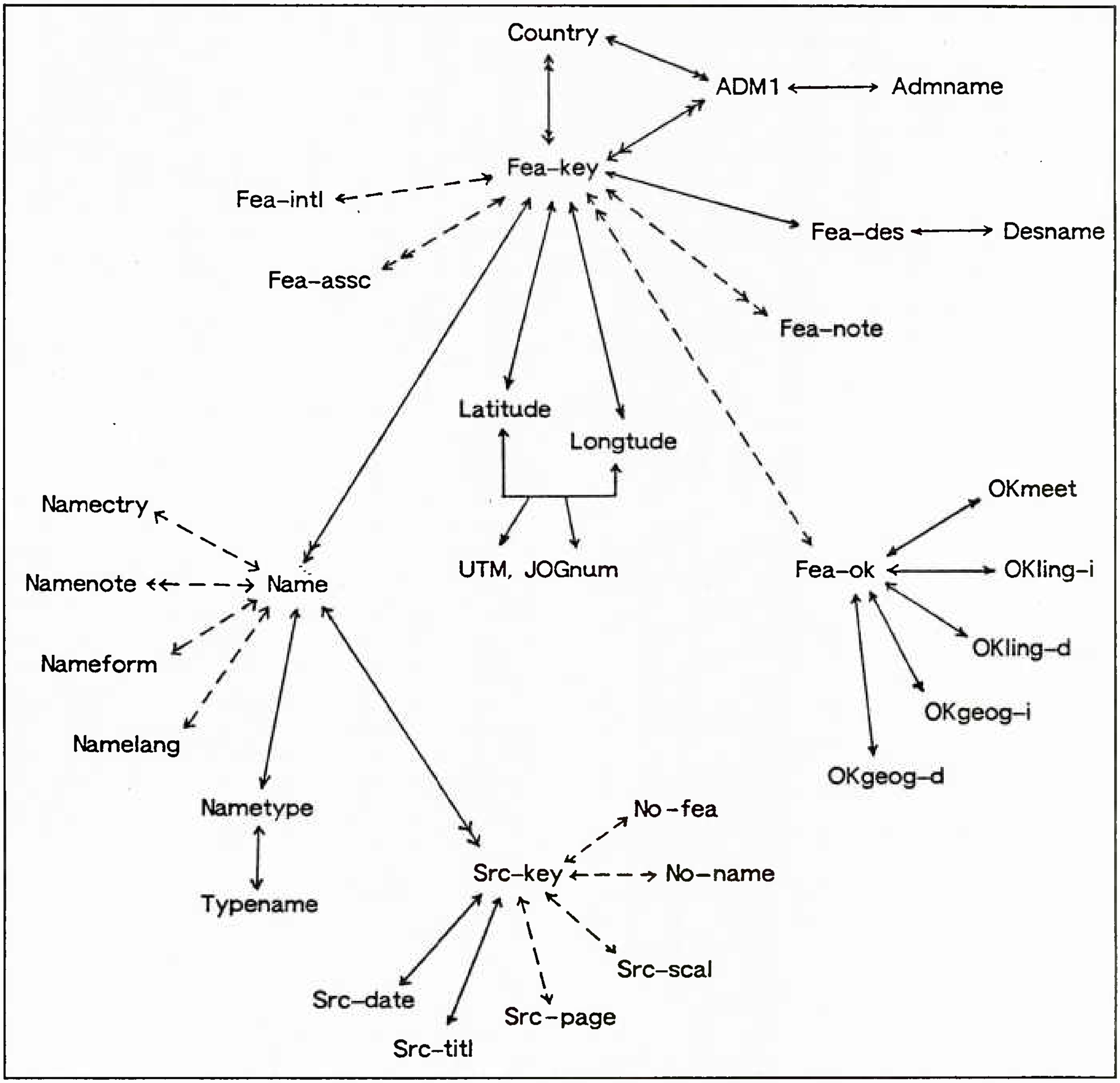


Figure 2-1. Conceptual data model. Solid connecting lines indicate required data entities; dotted lines indicate optional entities. Arrows indicate a one-to-one, one-to-many, or many-to-many relationship between the connected entities.

character sets. Good discussions of foreign text processing methods are available in Becker (1984) and IEEE (1985). Langran (1985b) has evaluated the available foreign text processing methods within the framework of DMA's needs.

To simplify the initial conversion to digital methods, only Roman alphabet text will be used. All non-Roman text will be transliterated prior to entry into the data base. The transliterated text will, however, contain a variety of diacritics, special characters, and special symbols that must be supported by a keystroking scheme and by hard- and softcopy output devices. Appendix A lists all characters beyond the 26-character Roman alphabet that the system must be capable of handling.

Table 2–3. Prototype data dictionary.

ADM1
The administrative area code is a standardized 4-character alphanumeric denoting the country and administrative division in which a feature falls (the FIPS country code is its first two digits). Features that do not cross international boundaries have only one area code (a special code is used for features that cross administrative divisions). Features that do cross international boundaries have one area code for each country they are associated with

Admname
The proper name of the administrative area in which a feature falls.

Country
The proper name of the country in which a feature falls. Each country has many ADM1s and features. Each ADM1 has only one country. Some features will have more than one country.

Desname
The full (decoded) name of a feature's designation.

Fea-assc
Association to another feature in the data base. A feature may be associated to zero or many other features. Features are associated to one another through action of an analyst.

Fea-des
The type of feature a name refers to (e.g., populated place, stream).

Fea-intl
An indicator that a feature crosses or forms an international boundary (e.g., that the feature is associated to more than one country).

Fea-key
A unique alphanumeric used to reference individual features. The alphanumeric may be acquired from an external source rather than assigned by the DBMS.

Fea-note
Research notes concerning a feature.

Fea-ok
A key to the entry that contains a feature's names approval data. The same names approval data can apply to many different features, since names are often approved in bulk lots.

JOGnum
The number of the JOG sheet upon which the feature would fall. JOGnum can be computed from the feature's latitude and longitude.

Latitude, Longitude
A feature's geographic coordinates. Line and area feature coordinates are one point only.

Name
The proper name of a known feature stored.

Namectry
The country in which an international feature's name is used.

Nameform
States whether a name was transliterated from its non-Roman form or acquired from a Roman source. Only names in languages using non-Roman characters must use this data element.

Namelang
The language in which a name is used. Namelang is for features in multilingual countries and those that cross international boundaries. Three-character language codes will be provided.

Namenote
Research notes concerning a name.

Nametype
The type of name (e.g., approved, variant). Each name has one or more approval types.

No-fea
States when a feature is not mentioned in a reference source.

No-name
States when a name is not mentioned in a reference source.

Okgeog-d
The date when an entry is approved by a geographer.

Okgeog-i
The initials of the geographer who approved an entry.

Okling-d
The date when an entry is approved by a linguist.

Okling-i
The initials of the linguist who approved an entry.

Okmeet
The number of the Foreign Names Committee meeting at which the entry was approved.

Src-date
The date of a reference source.

Src-key
A unique key to the reference source citation. Up to ten reference sources will be allowed for each name. One source may refer to many different names.

Src-page
The reference source page on which the name is mentioned.

Src-scal
The scale of a graphic reference source.

Src-titl
The title of a reference source.

Typename
The full (decoded) description of a name's toponymic type.

UTM
The feature's UTM coordinates, which can be computed from its latitude and longitude.

DATA CAPTURE

Most of the arduous job of populating the data base will be assumed by DMA personnel. The contractor will be responsible for loading an initial 5 million names (provided by DMA on 9-track tape) and for providing data capture. The GNPS will be populated using three major technologies: input from magnetic media, keystroked input, and input from a digitizing table. A fourth technology, optical character recognition, is of interest but will not be pursued for the GNPS. These four technologies are discussed below.

Input from magnetic media. In addition to DMA's tape archive of gazetteer names and names data, other countries and U.S. government agencies have digitized names. Thus, software to load formatted names data into the data base must be available.

Keystroked data entry. The most common, but also the most time-consuming, names entry method will be keystroking. Thus, input utilities for non-programmers are needed. At minimum, keystroking utilities should include a means of bulk-entering lists of names with attributes and a way to add attributes to names already in the data base. The bulk-entry utility should allow a non-programmer to define a file format (e.g., the data entities that will be entered) then keystroke lists of names with selected attributes, with or without system prompts. When data is added to a single name, it should be possible to view all the name's data currently stored.

Digitizing table entry. Positions for many of the new names added to the data base following GNPS delivery will be digitized from DMA maps and charts. Thus, GNPS software must be designed to expedite digitization.

Optical character recognition (OCR). The option of using OCR to capture names was evaluated and discarded. The FPNF cards are mostly handwritten; all FPNF diacritics are added by hand. These two factors make the use of OCR to capture FPNF data relatively unrewarding. Capturing map names via OCR could greatly aid the data base population process but the technology is not mature, making it too technically risky and expensive for this procurement. Possible map OCR methods include the use of a scanning wand (similar to those used in retailing) so an analyst can scan map names selectively. Alternatively, the entire map can be scanned and its names isolated and converted to ASCII.

MACHINE TRANSLATION

DMA's linguists currently provide DMA with translation capabilities and are likely to continue in that capacity for translations of idiomatically complex text. For rote translation of technical materials, however, a machine translation system could be very useful. The planned procurement does not include a translation system, but DMA may one day add such a capability.

Translation software, available today, should improve considerably as the technology matures. One company has developed an IBM PC-based translation system that is used by such corporations as ITT, Siemens, and Xerox (Dunn, 1982). The software is being implemented on a DEC VAX by its vendor, Weidner Communications, and is likely to become available on a range of other common computer systems.

PERSONNEL

Personnel needs will expand once the GNPS is introduced. The linguists and geographers employed today will remain the focal users of the new system. However, the computer skills of this specialized group are largely undeveloped. A computer staff must be brought in to tend to daily system and data base needs, to develop applications programs that will facilitate system use, and to assist the toponymists in using the system.

The Data Base Administrator (DBA) is the person who will dictate the GNPS's success or failure. The GNPS DBA should double as the system manager. He/she should control system and data access, mastermind improvements to system and data structuring, and supervise the computer staff.

Initially, the computer staff must include a minimum of two programmers, one for applications and one for system programming. A decrease in programming staff may be possible later. But during the critical first year, two full-time programmers are needed to ensure that the system is utilized fully with a minimum of down time, and to acquire the maximum amount of knowledge from the contractor during the one-year maintenance period following system delivery.

Also vital to system success is the establishment of an Information Center (IC). The IC staff consults with users when they have problems; arranges training; plans for future upgrades; proposes new utilities or changes to user interfaces; and facilitates system use through such devices as a newsletter or digital "bulletin board," documentation maintenance, and improvements to documentation. ICs defuse the frustrations that occur when new systems are introduced. By installing a facility whose function is communication, programmers can work with fewer interruptions from users needing assistance with the system, and the system's users waste less time wrestling with system use problems. ICs have repeatedly been proven to increase productivity and personnel satisfaction. One permanent IC staff member may be sufficient for a system of this size, particularly since the contractor must provide training. Temporary part-time assignments of toponymists to the IC could be helpful; the "sabbatical" allows a user to hone his skills and develop innovative ways to use and upgrade the system.

3. Technical answers to GNPS requirements

USER INVOLVEMENT IN GNPS REQUIREMENTS DEFINITION

The GNPS's ambitious objective is to convert a fully analog, highly idiosyncratic operation to a digital operation using standardized forms and methods. To define the scope of the GNPS, the future users' desires were weighed against current commercial capabilities in each technical area.

A group of five DMA linguists and geographers consulted extensively with NORDA and DMA technical and management personnel. A series of meetings were held with several weeks separating each. During the meetings, users described their information needs and current production and research activities. Between meetings, NORDA designed prototype interfaces for discussion and users assembled materials that were requested by NORDA.

Current names processing procedures are highly individualized; each analyst has preferred methods and nomenclature. Far from complicating the requirements analyst's task, however, dissent among analysts over procedure provoked extended discussions that were far more revealing than an orderly question-and-answer period would have been. In several cases, standards were agreed upon as a result of the discussions.

The next subsections describe prototype methods to manipulate and view data and special characters. The methods described below were developed iteratively, some by NORDA, some by DMA's names analysts. Their purpose was to ensure that the names analyst's needs were fully understood in a practical sense. Thus, when users were asked if a particular format would work, they had an opportunity to revise it until they were satisfied that it would. The prototyping was not intended to dictate the final GNPS's design; this is left to the contractor.

DATA MANIPULATION

At the first of several meetings, the future users answered a series of questions about their data manipulation needs. Responses were recorded by NORDA in the form shown in Figure 3–1.

Discussions concerning the questionnaire revealed many generalities and a few specifics that are helpful to GNPS design. Most important to understand: the GNPS will support an environment that is devoted equally to research and production. All data must be accessible in any combination; which information will be missing when seeking a particular record is impossible to predict. The names analysts were able to distinguish two levels of relative frequency for using a data element as selection criteria when accessing names data (Table 3–1).

Queries in support of production may be the least demanding upon the system. Since production schedules are planned months in advance, an analyst could submit a request to compile a subfile of all names data pertinent to features within a given country or rectangular area (e.g., map sheet) hours or even days in advance (the exception to this rule is during crises, when production needs become urgent; at such times, however, other system usage can be curtailed to improve response time). As long as the compiled subfile retains the data base relationships and can be worked in a query/response mode the analyst can remain within that limited data space, thereby greatly reducing his/her demands on the system. Similarly, production formats will be fixed. Standard software to support production can be developed as experience with the system is gained, and the DBA can chart navigation paths in the data base to speed access to frequently used data elements.

The most unpredictable queries, e.g., ad hoc requests for information, often require the quickest response. In such instances the individual who requests the information may know a name's pronunciation (but it won't necessarily be the name currently approved), the named feature's type (but it may be inaccurate), the names of some neighboring features and similarly limited information about them, and the feature's approximate position. Or, a request may be made for a listing of all approved names in a given region meeting certain criteria, using an output format previously unknown to the system.

Counting certain data elements will be an important GNPS capability (e.g., how many names meet certain criteria, how many variants are there for a given name). Sorting will also be an important GNPS feature. A standard method exists for sorting placenames, to which the GNPS must adhere (see Appendix B). Flexible ways to describe position are less important; users agreed to limit such descriptions to the coordinate corners of rectangular windows, stated absolutely or relatively. Thus, for the first implementation, other ways of describing positional criteria (e.g., via radial distance from a given point or location within a given map sheet's boundaries) are not needed.

SCREENS AND MENUS FOR THE USER INTERFACE

A number of standard formats and protocols are already in use among the GNPS's future users. The FPNF cards,

Queries to the names data base can be phrased in a number of ways. Rate how useful each query type shown below would be to your application. Rate how often you might query the data base in each way. Use the scales shown below.

Utility: 1-not needed 2-somewhat useful 3-useful 4-essential

Frequency: 1-less than once/month
2-several times/month
3-several times/week
4-several times/day

	Utility	Frequency	Batch ok
1. "Find all name within...."			
-map sheet X	______	______	______
-province/state X	______	______	______
-a geographic area defined by minimum and maximum latitude and longitude	______	______	______
-country X, within n (mi/kms) of point (X,Y)	______	______	______
-other	______	______	______

____
(area code)

____
(jog sheet)

2. "Find all names within a given area whose ______ (see list below) matches a given ______ (see list below)."

-feature code	______	______	______
-spelling	______	______	______
-feature class, e.g., hypsographic	______	______	______
hydrographic	______	______	______
cultural	______	______	______
vegetation	______	______	______
-review dates	______	______	______
-source	______	______	______
-FNC date	______	______	______

3. What other types of queries might be useful?

Figure 3-1. Working group response form.

Table 3–1. Relative frequency of accessing names data.

Frequently queried data entities: Country Placename Type of name Feature designation Latitude/longitude Administrative area code International feature indicator
Less frequently queried data entities: Association with other features Geographer's initials Date of geographic review Linguist's initials Date of linguistic review Foreign names committee meeting number Reference source date, title, and scale Form of name in reference source UTM position JOG sheet number Research notes Country of name (for international features) Language of name (for features in multilingual areas)

for example, have a spot for each standard data entry. Gazetteers are another well-understood names data format. Multiple names for a single feature are always shown on listings in a particular order. And, as mentioned previously, placenames are alphabetized by a precise set of rules. The working group unanimously agreed that the new system should comply with current protocols whenever feasible. A cooperative effort produced the following prototype screens, menus, and formats.

The Standard Query Response Format (SQUERF)

When an analyst is interactively searching the data base for names meeting a certain criteria, a standard response format is needed. The SQUERF (Fig. 3–2) displays summary information about a name in a columnar order acceptable to the analysts. SQUERF data are alphabetically sorted according to the standard system. Long SQUERF listings can be scrolled.

The SQUERF serves as a window into the data base. Analysts can move a cursor to any name on the SQUERF listing and summon a complete set of the data base information associated with that name in Report card form

Type	Name	ADM1	Desig	Lat	Long
A	Baile	EN06	PPL	29°36′N	42°44′E
A	Baile	EN06	STM	29°36′N	42°44′E
VI	Baile	EN07	STM	30°40′N	40°21′E

Figure 3–2. The SQUERF, which serves as a data menu.

(see next subsection). Thus, the SQUERF's primary purpose is as a data menu.

The Report Card

The Report Card (Fig. 3–3) displays detailed data concerning a single named feature. It is also an interactive device for adding to or modifying existing information, or adding a new named feature to a working file. The Report Card is an attempt to match the amenities that names analysts have enjoyed using the FPNF, namely, an ability to view all names information for a single feature and modify it on the spot, seeing the changes as they are made.

Country Name Area Code	Fea. Des. (Assoc.)	Latitude Longitude	UTM JOG
Type	Name		
Type	Name		
Type	Name (language, a 3-character abbreviation)		
Type	Name (country, a 2-character abbreviation)		
.	.		
.	.		
.	.		
FNC:		Geographer: MK, 1/87 Linguist: PP, 2/87	

Figure 3–3. The Report Card form.

The Report Card's top two lines provide information on the feature: its position, designation, the presence of associated features, and the administrative area and country in which it lies. Position is shown by degrees, minutes, and seconds; UTM coordinates; and JOG sheet number. If the feature crosses an international border "International" is written in place of a country name. "Assoc" is shown only when other features stored in the data base have been logically associated to the Report Card feature.

The next 15 lines contain fields for the type of name (i.e., local official, conventional) and the name itself. A name with a language related to it is followed by a 3-character language code in parentheses (i.e., Lac Erie (FRN)). International feature names are followed by a 2-character country abbreviation in parentheses. Because more then 15 names/feature can occur, this Report Card section can be scrolled independently of the rest of the Report Card.

The next two Report Card lines hold approval data: the FNC meeting number, the linguist's review date and initials, and the geographer's review date and initials.

The bottom section of the Report Card is used interactively by the analyst to request source citation listings (Fig. 3–4a), research notes for features or names (Fig. 3–4b), international feature information (Fig. 3–4c), and information regarding other features that have been logically associated to the Report Card's feature (Fig. 3–4d).

Names are output to the Report Card in the following order, which is considered standard by DMA's names analysts.

- If both long and short forms (local official or conventional) of a name are listed, the short form always precedes the long form.

- If the named feature is located wholly within one country, the local official name(s) precede(s) the conventional name, e.g., Bruxelles [French]; Brussel [Flemish]; Brussels [conventional]. The same sequence applies to names for administrative entities, e.g., HaMerkaz, Mehoz [Hebrew], Central District [conventional].

- If the named feature is located in more than one country, the conventional name precedes the local official name(s), e.g., Drava River [conventional]; Drava [YUGOSLAVIA] Drava [HUNGARY]; Drau [AUSTRIA]. The same sequence applies to country names, i.e., Finland, Republic of [conventional]; Suomi [Finnish short form]; Suomen Tasavalta [Finnish long form]; Republiken Finland [Swedish].

- Variant names follow approved names.
- Not verified names appear last.

To add data to static Report Card sections the cursor is moved to the applicable field, an "Insert" function key is struck, and the new information is typed. To alter or delete Report Card data the cursor is moved to the space occupied by the erroneous information and an "Edit" or "Delete" function key is struck. To input coordinates from the digitizing table the analyst moves the cursor to the "position" field, digitizes the position from a map fastened to the table, and strikes a function key. If any of the position fields are altered (e.g., geographics, UTM coordinates, or JOG sheet number) the system recomputes the other two position fields. Editing the Report Card's dynamic lower section is done by summoning the information to be edited to that section and altering it as described previously.

Hardcopy Report Card facsimiles can be produced on the printer. Additions, modifications, and deletions appear immediately on the Report Card but are not made to the

Country Name Area Code	Fea. Des. (Assoc.)	Latitude Longitude	UTM JOG
Type Name Type *Name* Type Name Type Name			
FNC:		Geographer: MK, 1/87 Linguist: PP, 2/87	
Date Source title (R) Date Source title (N) Date Source title (R)			

Figure 3–4a. Listing source citations on the Report Card. Dates, titles, and Romanization information is listed for the name indicated by the analyst's cursor (second name on the list). "R" means the name appears in its Romanized form, "N" means the name appears in its non-Romanized form. It is possible to scroll or page independently through the source information if its length exceeds that allotted.

Country Name Area Code	Fea. Des. (Assoc.)	Latitude Longitude	UTM JOG
Type Name Type *Name* Type Name Type Name			
FNC:		Geographer: MK, 1/87 Linguist: PP, 2/87	
Research note Research note Research note			

Figure 3–4b. Listing research notes on the Report Card. All stored research notes are listed for the name indicated by the cursor. It is possible to scroll or page the research notes independently of the top half of the form if their length exceeds the space allotted.

International Area Code	Fea. Des. (Assoc.)	Latitude Longitude	UTM JOG
Type	Name		
Type	Name		
Type	Name (SPN)		
Type	Name (FR)		
.	.		
.	.		
.	.		
FNC:		Geographer: MK, 1/87 Linguist: PP, 2/87	
Borders crossed by feature:			
Country			
Country			

Figure 3-4c. Listing international feature names on the Report Card. The analyst moves his cursor to the country field, which reads "International." All countries in which that feature lies are listed in the lower half of the Report Card.

master data base until such entry has been formally requested. Instead, they are stored in a temporary working file that the analyst may opt to discard, save, or merge with the data base following his Report Card session. The work of analysts without read/write data base privileges is written to a file for later processing by a privileged analyst. In all cases, writing to the data base is a batch process.

Report Cards can also be used to add new features to the data base. To do this, the analyst summons a blank Report Card to the workstation screen. Data is written on the Report Card by moving the cursor to a field and typing the information using the same utilities described in the previous paragraphs.

The Query Form

The Query Form is a standard device for initiating data base queries. The DBMS is expected to have a query language, but many names analysts will prefer to use a menu-based approach to submit their requests for information.

To initiate a query, a Query Form (Fig. 3-5) is summoned to the workstation screen and the first field is toggled to read "Criteria." Steps for query formulation and submission are pictured in Figures 3-6a to 3-6g. The query form is forgiving; analysts can move forward or backward in a stepwise fashion if errors or omissions are discovered.

Country Name Area Code	Fea. Des. (Assoc.)	Latitude Longitude	UTM JOG
Type	Name		
Type	Name		
Type	Name		
Type	Name		
(etc.)	(etc.)		
.	.		
.	.		
.	.		
FNC:		Geographer: MK, 1/87 Linguist: PP, 2/87	
Associated Features:			
Name	Feature designation	Position	
Name	Feature designation	Position	

Figure 3-4d. Listing associated feature data on the Report Card. The analyst moves his cursor to the "Assoc" field to list associated feature data in the lower half of the Report Card. It must be possible to summon an associated feature's Report Card from the current Report Card, and return to the current Report Card without any intervening queries.

Criteria/Response (toggle)					
Name	Feature Designator	Latitude	Longitude	UTM	JOG
Country	ADM1	Type	Internat'l	Associated	
Source					
Source date	Roman/Non-Roman				
FNC	Geographer	Geographer Date			
	Linguist	Linguist Date			

Figure 3-5. Query Form. The analyst indicates whether the form describes query criteria or a query response format by toggling the top field.

The default output form for queries is the SQUERF. Alternate output needs are described via the Query Response Form. To define an alternate output form, the

```
Criteria

       Feature
Name   Designator  Latitude      Longitude      UTM  JOG
B#L#   PPL         23°45'S+2-2   25°44'E+2-2
       STR

Country    ADM1    Type    Internat'l    Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer    Geographer Date

               Linguist      Linguist Date
```

Figure 3-6a. Query criteria are added beneath the fields shown on the Query Form. In this example, qualifying names begin with a B, contain an L, and name populated places or streams within 2 minutes of the stated latitude and longitude.

```
Criteria

        Feature
Name    Designator    Latitude    Longitude    UTM    JOG
                      23°45'S     25°44'E
                      24°45'S     24°44'E

Country    ADM1    Type    Internat'l    Associated
                   A       Y
                   C#
Source

Source date    Roman/Non-Roman
>1982              R

FNC            Geographer    Geographer Date

               Linguist      Linguist Date
```

Figure 3-6b. This example shows criteria for all approved and conventional names of features that cross an international boundary and lie within a given latitude/longitude window, that also appear in their Romanized form in sources more recent than 1982—an eccentric but not impossible query.

user summons a Query Form to the screen, toggles the first field to read "Report," and works with the form as illustrated in Figures 3-7a to 3-7c.

```
Criteria      Report: SQUERF    Gazetteer    Stored    Alternate

       Feature
Name   Designator  Latitude      Longitude      UTM  JOG
B#L#   PPL         23°45'S+2-2   25°44'E+2-2
       STR

Country    ADM1    Type    Internat'l    Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer    Geographer Date

               Linguist      Linguist Date
```

Figure 3-6c. When the analyst signals that query criteria are complete, the system prompts for the query response format. The default response format is the SQUERF. Designing alternate query response forms is discussed in the next subsection.

```
Criteria      Report: SQUERF

       Feature
Name   Designator  Latitude      Longitude      UTM  JOG
B#L#   PPL         23°45'S+2-2   25°44'E+2-2
       STR

Country    ADM1    Type    Internat'l    Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer    Geographer Date

               Linguist      Linguist Date

Batch/Interactive
```

Figure 3-6d. The system prompts the user for batch or interactive query submission.

MANIPULATING SPECIAL CHARACTERS

Keyboard input

Many lively discussions occurred over GNPS keyboard design but no consensus was reached. This section describes some of the alternatives that were discussed; each has proponents at DMA.

```
Criteria        Report: SQUERF

       Feature
Name  Designator  Latitude        Longitude      UTM  JOG
B#L#  PPL         23°45'S+2-2  25°44'E+2-2
      STR

Country   ADM1   Type   Internat'l   Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer   Geographer Date

               Linguist     Linguist Date

Batch                                    Priority: 2   3   4
Output: File/Printer
```

Figure 3–6e. If the user selects batch submission, the system prompts for priority and output destination. Priority defaults to 4, output destination defaults to file.

```
Criteria        Report: SQUERF

       Feature
Name  Designator  Latitude        Longitude      UTM  JOG
B#L#  PPL         23°45'S+2-2  25°44'E+2-2
      STR

Country   ADM1   Type   Internat'l   Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer   Geographer Date

               Linguist     Linguist Date

Batch                                    Priority: 4
Filename: Myfile
```

Figure 3–6f. If the user elects to send the query response to a file, the system prompts for filename.

Funding constraints prevent special tooling of workstation keyboards. Therefore, the difficulty of designing the GNPS keyboard lies in fitting approximately 32 extra symbols onto an off-the-shelf keyboard (the number of extra symbols might vary with the implementation strategy; more than 32 might be required).

An ETL prototyping effort produced a system where the world's linguistic systems were divided into regional

```
Criteria        Report: SQUERF

       Feature
Name  Designator  Latitude        Longitude      UTM  JOG
B#L#  PPL         23°45'S+2-2  25°44'E+2-2
      STR

Country   ADM1   Type   Internat'l   Associated

Source

Source date    Roman/Non-Roman

FNC            Geographer   Geographer Date

               Linguist     Linguist Date

Interactive
Output: Terminal/File/Printer
```

Figure 3–6g. If the query is submitted interactively, priority defaults to 1. Terminal, printer, or file output must be selected. Default interactive output is to terminal.

```
Response

        Feature
Name    Designator    Latitude    Longitude

Country   ADM1
S         L

FNC
```

Figure 3–7a. Specifying an alternate query response form. The analyst toggles the "Criteria" field to "Response," then deletes all but the desired output fields. In this example, output fields include the name, feature designation, geographic coordinates, country, area code, and approval code. The country name is abbreviated (S) and the area code is spelled out (L). The list will be formatted by the system into a table and alphabetized using the protocol described in Appendix I.

diacritics sets (REDS), allowing all the diacritics and special symbols for a particular language to be accessed from a

Response

Feature
Designator
U

Figure 3–7b. In this example, the response will be a list of all unique feature designators in the qualifying data set. It is also possible to count all unique feature designations in the qualifying data set.

Response
Diacritics/No diacritics (toggle) (if no action is taken, diacritics are included).

Name | Feature Designator | Latitude | Longitude

Country | ADM1

FNC

Sorting: Standard/User-specified
Store under what name? (if no name is entered, the form is not stored)

Figure 3–7c. When the response form is completed, the system prompts for a name under which to store the form and asks whether the output should contain diacritics (default is no diacritics). The user is also given the option of specifying the sorting method; sorting can be performed on any field.

regular QWERTY keyboard with minimal use of function keys. Such a system of partitioned access to diacritics, etc., has merit, but the analyst should be able to see what character(s) are assigned to what key under the current keyboard configuration. Some suggested methods of labeling keyboards whose character/key assignments vary are: displaying the keyboard on the monitor screen (on analyst request), having different physical keyboards, and having keyboard templates.

Other diacritics input ideas suggested accessing diacritics without using the standard keyboard. Alternatives include using an adjunct keyboard, an adjunct digitizing tablet equipped with template and pointing device, or a windowed template on the workstation screen with a pointing device. None of the names analysts in the working group claimed to be touch typists; however, this method would prevent any names analyst using diacritics and special symbols from ever becoming a touch typist.

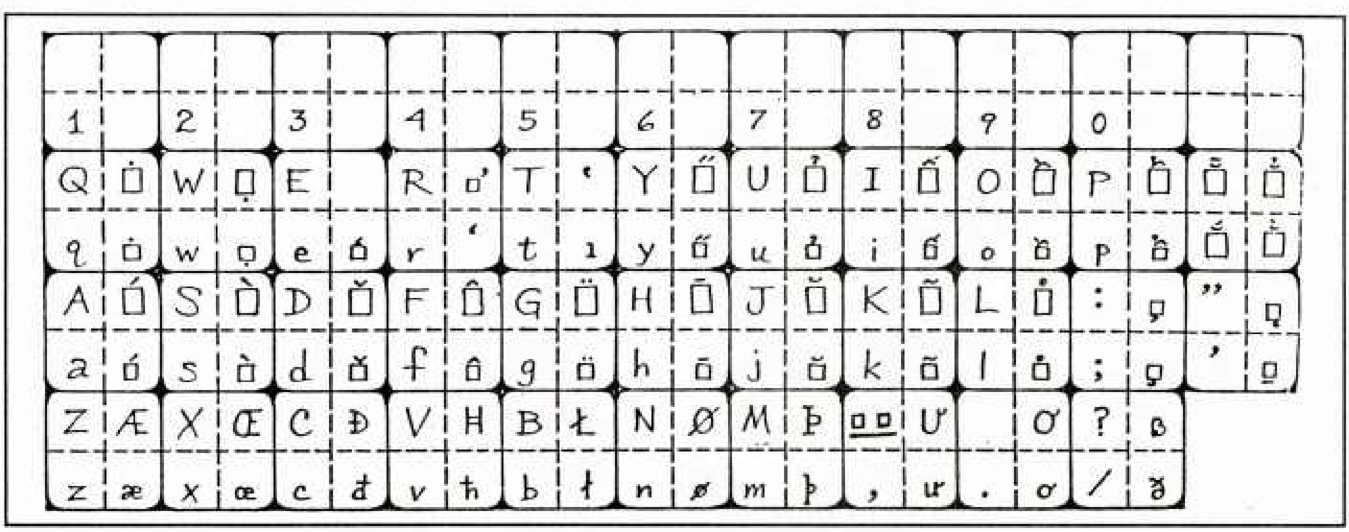


Figure 3–8. Prototype GNPS keyboard.

Several of the names analysts designed a system of transliterated text input that fits all necessary characters onto a standard typewriter keyboard (Fig. 3-8). To use this system, characters in the SW and NW quadrant (e.g., the standard keyboard characters) are produced using standard typing practices: striking the key alone produces the SW character, striking the key in conjunction with the shift key produces the NW character.

Characters in each key's SE quadrant are typed by striking an "alternate character" function key in conjunction with the character key. Characters in NE quadrants are typed by striking an "alternate upper case character" function key in concert with the character key.

Coding diacritics and special symbols

The diacritics keystroking order that is most in harmony with the way diacritics are perceived is to type the character first, followed by all its diacritics typed in no particular order. This input order has been stated by DMA as a GNPS requirement. Although necessary, it incurs several complications, discussed below.

NAPLPS codes. The North American Presentation Level Protocol Syntax (NAPLPS), backed by the American National Standards Institute, has developed a code for diacritics and other foreign symbols that fits within the 8-bit ASCII protocol (ANSI, 1983) (see Appendix C). Codes are available for diacritics alone and diacritics attached to characters. However, DMA linguists state that the NAPLPS list contains diacritics/character combinations of dubious validity and lacks some that are needed, which makes rote implementation of NAPLPS undesirable. Adequate means of expanding on the NAPLPS diacritics set do exist; thus, use of the NAPLPS standard to the extent possible is a GNPS requirement.

Standard ANSI/ASCII practice is to follow every printed character with a "character feed," e.g., the protocol moves the printhead forward one space to avoid overprinting. The NAPLPS protocol defines diacritics as nonspacing characters, e.g., character feeds do not follow diacritics. Thus, the NAPLPS storage protocol requires that diacritics precede their characters, in direct opposition to the keystroking order of future GNPS users. This implies that GNPS software must invert character/diacritics combinations for permanent storage, a clumsy but not difficult endeavor. Sufficient importance was attached to both analyst comfort and adhering to NAPLPS to justify the additional step.

Appearance of printed text. Two different printing standards exist for text with diacritics. Diacritics text on the analyst's workstation monitor is permitted some latitude in its appearance. It is likely that a character and its diacritics will be assembled within the character's dot matrix in real time. Since a single diacritic can be applied to many characters and one character can host many different diacritics, sometimes in groups, exact placement of diacritics relative to characters may be a problem on the workstation screen. For example, an acute accent may be placed over a's, c's, e's, g's, etc. A circumflex or breve may occur in tandem with the acute accent over certain letters in certain languages. Thus, although a diacritic is expected to appear in correct relation to its letter and be easily recognized, perfection is not expected for workstation displays.

Conversely, typeset-quality diacritics text is expected on the printer. To achieve this, all possible diacritics/character combinations must be precomposed and stored as separate bit maps by the system. Thus, some further manipulation of character codes may be required at printing time to maximize efficiency.

DIGITIZING NAMES FROM MAPS

Maps are useful data sources for names, coordinate positions, feature designations, and reference sources (e.g., the map itself). Any possible way of expediting data capture should be incorporated into the GNPS, given the volume of data to be captured. A possible map data capture scenario follows.

Preparation

- Register the map to the digitizing table. Enter the scale, projection, and ellipsoid from the console keyboard. This information is used by software to transform digitizing table coordinates to geographic coordinates.
- Enter the map title and date from the console keyboard. Define whether map names are in Roman or non-Roman form. This information is the default reference source for all names entered during the digitizing session.
- Select a feature designation template and fasten it to an active part of the digitizing table that is not occupied by the map. The template is comprised of ½-inch cells, one for each of the 25 most common feature designations. If multiple feature designation templates are available (e.g., for several physical environments) the software must be notified of which template is being used.

Digitization

- Digitize the named feature's location. Type the placename.

- The map is added to the record as a reference source by default. Other reference sources can be added from the keyboard if necessary. The analyst can summon to the workstation screen a listing of that country's reference sources and indicate those that apply with a cursor.
- If the feature designation is contained on the template the cursor is moved to that location and the cell is digitized; otherwise, the feature designation is typed from the workstation keyboard.

MARKET SURVEY

The choice of GNPS hardware and software is left to the contractor. The market survey served mainly to constrain this procurement's scope so its cost would not exceed available resources and its performance specifications would remain within reach of off-the-shelf equipment.

GNPS configuration

A mainframe computer is the most proven way to handle large data bases and a number of proven DBMSs are available for such devices. But a quick look at the marketplace revealed that the price of a mainframe computer with peripherals and software is double-to-triple the entire GNPS budget. This problem, and the difficulty of constructing the physical environment required by such a computer system, seemed to obviate a mainframe-based configuration. Alternate solutions were investigated.

GNPS requirements were subdivided into initial requirements and optional upgrades. The minimum acceptable GNPS data base size is 20 million names and their associated data (although only about 5 million names with partially complete data will be ready to load in the first year). Twenty workstations, five digitizing tables, two 9-track tape drives, and two high-speed printers are needed in addition to the processors and mass storage. Software is needed to manage the data base, word processing, and systems functions. Contractors were asked to state the per/workstation, per/digitizing station, and per/500,000-name costs of upgrading the system beyond those minimum requirements so DMA could consider the option of increasing the GNPS's size.

Decentralized processing strategies were investigated. Inexpensive but powerful microcomputers have recently become widely available. Some notable examples are the Microvax and the Sun Workstation. Both are supported by a number of second-party software vendors selling data bases capable of supporting the GNPS. Configuration options abound; this class of microcomputer is powerful enough to host 3–5 terminals but inexpensive enough to serve as a workstation. One possibility is to have one such computer manage the central storage media and have less powerful models with local storage serving as individual workstations.

Advantages of the decentralized processing strategy are its expandability and backup capabilities. Disadvantages include the difficulty of designing and managing the system, a possible need to partition the names data base, and the fact that decentralized data base technology is not mature.

Monitor characteristics

Monitor resolution and viewer fatigue will have a major impact on GNPS user satisfaction. Flicker-free workstation monitors were a stated GNPS requirement. Resolution is a practical and esthetic matter. Because of the large dot matrix required to create character/diacritic combinations, poor resolution means that only small amounts of text can be viewed, and that characters and spacing will appear oddly unbalanced or too large.

The Smalltalk programming environment, though not a GNPS requirement, is an ideal GNPS capability. Smalltalk, a Xerox trademark, is used by the Xerox Star, Symbolics Lisp Processors, Sun Workstations, Audre Computers, and DEC's VAXstations, among others. The hardware specifications of those systems were compared and the minimum value for each monitor attribute was made a GNPS requirement.

Smalltalk's multi-window protocol is ideal for the comparisons, research, and cross-referencing that predominate in names processing. An analyst can summon a gazetteer file in one part of his/her screen and the feature designation definitions in another, or review a screenload of report cards of features associated to one another, or compare two files side by side. A second important Smalltalk asset is its mouse-based, user-friendly operation that helps analysts to focus on their work rather than on the discomforts and disadvantages of an unpleasant system.

4. Summary

This report described the needs of a digital geographic names processing system for DMA. Special attention was given to describing the available criteria for the decisions that needed to be made. By documenting the requirements analysis process, future users and managers of the GNPS will understand how and why their system came to be configured to its final form.

Attention was also given to describing the prototype forms and procedures that were developed over a series

of meetings between DMAHTC names analysts and the author. The prototypes illustrate ways to meet the needs of DMA's names analysts, having been examined and critiqued by the GNPS's future users. Use of the prototype methods is not a GNPS requirement.

At the time of publication, a contract award for work on the GNPS was scheduled to occur during the first quarter of FY 87, with initial system delivery and training occurring during the first quarter of FY 88.

5. Bibliography

American National Standards Institute (1983). *Videotex/Teletext Presentation Level Protocol Syntax.* Washington, D.C., X3.110-1983.

Becker, Joseph D. (1984). *Multilingual Word Processing.* Scientific American, v. 251, n. 1, July, pp. 96–107.

Brown, R., A. Zeid, A. Barnes, and E. Gough (1983). *Advanced Type Placement and Geographic Names Data Base.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Technical Note 189, January.

DMAHTC (1978). *Applications Branch, Geographic Names Data Base Division, Scientific Data Department.* Unpublished paper, Washington, D.C.

Dunn, John M. (1982). *Babel-matics.* Europe, November–December, pp. 28–29.

IEEE (1985). Computer. v. 18, n. 1.

Langran, Gail (1985a). *Geonames Processing System Functional Design Specification: Advanced Type Placement.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 100, March.

Langran, Gail (1985b). *Geonames Processing System Functional Design Specification: Advanced Symbol Processing.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 101, March.

Langran, Gail (1985c). *Geonames Processing System Functional Design Specification: Performance Specification.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 102, March.

Langran, Gail (1985d). *Geonames Processing System Synopsis.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 125, September.

Langran, Gail, Anne Downs, Barry Glick, and Warren Schmidt (1985a). *Geonames Processing System Functional Design Specification: Automated Alphanumeric Data Entry System.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 98, March.

Langran, Gail, Allen Barnes, and Steven Miller (1985b). *Geonames Processing System Functional Design Specification: Geographic Names Data Base.* Naval Ocean Research and Development Activity, NSTL, Mississippi, NORDA Report 99, March.

U.S. Army Engineer Topographic Laboratories (1983). *Names Type File System.* A consulting report prepared by Electromagnetic Compatibility Analysis Center, Annapolis, Maryland. April.

Appendix A: Diacritics and special character requirements

The special font set necessary to accommodate the diacritical marks in the data system should be a floating set, i.e., it should be possible to position any diacritic above or below an upper- or lowercase letter (or ligature) for display and printing purposes.

Diacritics:

á	(acute accent)
è	(grave accent)
ü	(dieresis)
ā	(macron)
ô	(circumflex)
š	(wedge)
ñ	(tilde)
å	(circle above)
ŭ	(breve)
ż	(dot above)
ő	(double acute accent)
g̓	(inverted apostrophe above)
ẻ	(pseudo question mark)
ç	(cedilla)
ḏ	(sub-macron)
s͟h	(extended sub-macron)
ṣ	(sub-dot)
ķ	(sub-comma)
ę	(right hook)

Double diacritics:

ḯ

ñ́

ấ

ầ

ẫ

ẩ

ậ [This entry may prove unnecessary since the circumflex and sub-dot are already used singly — if they can be positioned in two successive steps, one above and one below the letter.]

ắ

ằ

ẵ

ẳ

ặ [This entry may prove unnecessary since the breve and sub-dot are already used singly — if they can be positioned in two successive steps, one above and one below the letter.]

Special letters and symbols:

æ (Æ)

œ (Œ)

ı (I)

ø (Ø)

ł (Ł)

đ (Đ)

ð

ƿ (Ƿ)

ħ (Ħ)

ß

t' (half-space apostrophe)

d' (half-space apostrophe)

ŀ' (half-space apostrophe)

the half-space apostrophe goes with only these three letters

ư

ơ

ʼ (apostrophe)

ʼʼ (double apostrophe)

ʻ (inverted apostrophe)

Appendix B: Alphabetization rules

This section is for alphabetizing Foreign Place Names Information File (FPNIF) entries according to the alphabetization rules approved by the Board on Geographic Names. The rules are applied sequentially, and where there is seeming conflict, each rule takes precedence over application of the rule which follows.

(1) Alphabetize letter by letter to the comma, disregarding spaces, hyphens, and periods.

(2) If the letters of the names preceding the comma are the same, put in order by words following the comma:

Newark, Mount
Newark, River
Newark, Station
Newark Lake
Newark Lake, Mount
Newark Station

(3) If all the letters of the name are the same, put in order by hyphenation and spacing:

(a) Two-word form: New Ark

(b) Hyphenated form: New-Ark

(c) One-word form: Newark

(5) Names identical in all ways will be put in order by coded designation:

(a) Sovereignties, administrative division, in this order: PCLI, ADM1, ADM2, etc. Exceptions to this exist in some areas. These are noted as follows:

1. As Suwayda [short form]; Muhafazat as Suwayda [long form] (ADM1)

   As Suwayda [short form]; Mintaqat as Suwayda [long form] ADM2

   As Suwayda, PPL.

2. Dar a, PPL.
   Dar a, Mintqat, ADM2
   Dar a, Muhafazat, ADM1

(b) Populated places and qualified populated places: treat PPL, PPLQ, PPLX as a single designation and interfile by coordinates as indicated in Section j(6).

(c) All other features, in order by coded designation (AIRF, BAY, CAPE, DAM, DSRT, MT, RSTN, STM/STMI/STMX (treat these last three as a single designation and interfile by coordinates as indicated in Section j(6).

(6) Entries with identical names and designations are sorted by coordinates:

(a) Lowest to highest latitude.

(b) Lowest to highest longitude figure, if latitude is the same.

(7) Geographic names containing spelled-out numbers will be filed in alphabetical order. Examples:

Fourth of July, Mount	Hillsboro
Hilldale	Hill Ten
Hill Five	Hill Two
Hillmont	Seven Mile Creek
Hill Number One	

(8) Geographic names containing Arabic or Roman numerals in an initial, internal, or final position will be filed as follows:

(a) Names containing FINAL numbers will be filed following names containing the same sequence of letters without a final number. They will follow any names of the same spelling with reversed generics, and precede any names of the same spelling followed by "a." (Examples 7-14) Arabic-numeral names, in numerical order, will precede Roman-numeral names in each of two categories: names without reversed generics (Examples 7-10); and names with reversed generics (Examples 11-14).

(b) Names containing INTERNAL numbers will be filed following names described in (a), in numerical order, Arabic numerals preceding Roman numerals in each category, as noted above. (Examples 18, 19, 34, 35)

(c) Names containing INITIAL numbers, consisting solely of a number, or of a number followed by a comma and generic term, will be filed in numerical order following "z" names or their equivalent in file. (Examples 37-43)

A complete numerical sequence of names containing Arabic numerals will precede those containing Roman numerals. Names containing initial numbers followed by letters will follow names described above, in numerical order, with all Arabic-numeral names preceding all Roman-numeral names. (Examples 44-50)

Alphabetic order of letters following numbers but preceding commas will be used as the second criterion in arranging within a specific Arabic- or Roman-numeral group. (Examples 45, 46)

Alphabetic order of generics will be considered only when duplicates of a particular Arabic- or Roman-numeral specific name are encountered. (Examples 47, 48)

Examples:

1 Abou Kous
2 Abou Kous I
3 Abou Kous II
4 Abou Kous Djekebe
5 Bang Bo

6 Bang Bo, Amphoe
7 Bang Bo 1
8 Bang Bo 2
9 Bang Bo I
10 Bang Bo IV

11 Bang Bo 2, Amphoe
12 Bang Bo 4, Khao
13 Bang Bo I, Khao
14 Bang Bo IV, Ao
15 Bang Boa

16 Bang Bo Land
17 Bang Bo Land I
18 Bang Bo I Land
19 Bang Bo I Land, Mount
20 Bang Bon

21 Mariposa
22 Mariposa, Arroyo
23 Mariposa 2, Sierra
24 Mariposa II, Arroyo
25 Mariposa Abajo

26 Mariposas
27 Mariposas, Laguna
28 Mariposas 44, Desvio
29 Mariposas I
30 Mariposas I, Ventisquero

31 Mariposas Viejo
32 Seccion de Julio
33 Seccion de Julio 1
34 Seccion 9 de Julio
35 Seccion 10 de Julio, Parada

36 Seccion Delfino
37 483
38 1776
39 XVIII
40 1, Canal

41 2, Laguna
42 II, Laguna
43 XXIII, Canal
44 1 de Abril
45 1 de Abril, Estancia

46 1 de Diciembre
47 2 Amigos, Arroyo
48 2 Amigos, Cerro
49 4th of July River
50 IX de Julio, Ensenada

Appendix C: NAPLPS character and diacritics codes[1]

Graphic	Name or Description	Coded Representation
a	lower case a	6/1
A	upper case A	4/1
á	lower case a with acute accent	S 4/2 6/1
Á	upper case A with acute accent	S 4/2 4/1
à	lower case a with grave accent	S 4/1 6/1
À	upper case A with grave accent	S 4/1 4/1
â	lower case a with circumflex accent	S 4/3 6/1
Â	upper case A with circumflex accent	S 4/3 4/1
ä	lower case a with diaeresis or umlaut mark	S 4/8 6/1
Ä	upper case A with diaeresis or umlaut mark	S 4/8 4/1
ã	lower case a with tilde	S 4/4 6/1
Ã	upper case A with tilde	S 4/4 4/1
ă	lower case a with breve	S 4/6 6/1
Ă	upper case A with breve	S 4/6 4/1
å	lower case a with ring	S 4/10 6/1
Å	upper case A with ring	S 4/10 4/1
ā	lower case a with macron	S 4/5 6/1
Ā	upper case A with macron	S 4/5 4/1
ą	lower case a with ogonek	S 4/14 6/1
Ą	upper case A with ogonek	S 4/14 4/1

Graphic	Name or Description	Coded Representation
æ	lower case æ dipthong	S 7/1
Æ	upper case Æ dipthong	S 6/1
b	lower case b	6/2
B	upper case B	4/2
c	lower case c	6/3
C	upper case C	4/3
ć	lower case c with acute accent	S 4/2 6/3
Ć	upper case C with acute accent	S 4/2 4/3
ĉ	lower case c with circumflex accent	S 4/3 6/3
Ĉ	upper case C with circumflex accent	S 4/3 4/3
č	lower case c with caron	S 4/15 6/3
Č	upper case C with caron	S 4/15 4/3
ċ	lower case c with dot	S 4/7 6/3
Ċ	upper case C with dot	S 4/7 4/3
ç	lower case c with cedilla	S 4/11 6/3
Ç	upper case C with cedilla	S 4/11 4/3
d	lower case d	6/4
D	upper case D	4/4
ď or d'	lower case d with caron	S 4/15 6/4
Ď	upper case D with caron	S 4/15 4/4

(Continued)

Graphic	Name or Description	Coded Representation
đ	lower case d with stroke	S 7/2
Đ	upper case D with stroke, Icelandic eth	S 6/2
ð	lower case eth, Icelandic	S 7/3
e	lower case e	6/5
E	upper case E	4/5
é	lower case e with acute accent	S 4/2 6/5
É	upper case E with acute accent	S 4/2 4/5
è	lower case e with grave accent	S 4/1 6/5
È	upper case E with grave accent	S 4/1 4/5
ê	lower case e with circumflex accent	S 4/3 6/5
Ê	upper case E with circumflex accent	S 4/3 4/5
ë	lower case e with diaeresis or umlaut mark	S 4/8 6/5
Ë	upper case E with diaeresis or umlaut mark	S 4/8 4/5
ě	lower case e with caron	S 4/15 6/5
Ě	upper case E with caron	S 4/15 4/5
ė	lower case e with dot	S 4/7 6/5
Ė	upper case E with dot	S 4/7 4/5
ē	lower case e with macron	S 4/5 6/5
Ē	upper case E with macron	S 4/5 5/5
ę	lower case e with ogonek	S 4/14 6/5
Ę	upper case E with ogonek	S 4/14 4/5

(Continued)

Graphic	Name or Description	Coded Representation
f	lower case f	6/6
F	upper case F	4/6
g	lower case g	6/7
G	upper case G	4/7
ǵ	lower case g with acute accent	S 4/2 6/7
ĝ	lower case g with circumflex accent	S 4/3 6/7
Ĝ	upper case G with circumflex accent	S 4/3 4/7
ğ	lower case g with breve	S 4/6 6/7
Ğ	upper case G with breve	S 4/6 4/7
ġ	lower case g with dot	S 4/7 6/7
Ġ	upper case G with dot	S 4/7 4/7
Ģ	upper case G with cedilla	S 4/11 4/7
h	lower case h	6/8
H	upper case H	4/8
ĥ	lower case h with circumflex accent	S 4/3 6/8
Ĥ	upper case H with circumflex accent	S 4/3 4/8
ħ	lower case h with stroke	S 7/4
Ħ	upper case H with stroke	S 6/4
i	lower case i	6/9
I	upper case I	4/9
í	lower case i with acute accent	S 4/2 6/9
Í	upper case I with acute accent	S 4/2 4/9

(Continued)

Graphic	Name or Description	Coded Representation
ì	lower case i with grave accent	S 4/1 6/9
Ì	upper case I with grave accent	S 4/1 4/9
î	lower case i with circumflex accent	S 4/3 6/9
Î	upper case I with circumflex accent	S 4/3 4/9
ï	lower case i with diaeresis or umlaut mark	S 4/8 6/9
Ï	upper case I with diaeresis or umlaut mark	S 4/8 4/9
ĩ	lower case i with tilde	S 4/4 6/9
Ĩ	upper case I with tilde	S 4/4 4/9
İ	upper case I with dot	S 4/7 4/9
ī	lower case i with macron	S 4/5 6/9
Ī	upper case I with macron	S 4/5 4/9
į	lower case i with ogonek	S 4/14 6/9
Į	upper case I with ogonek	S 4/14 4/9
ĳ	lower case ij ligature	S 7/6
Ĳ	upper case IJ ligature	S 6/6
ı	lower case i without dot	S 7/5
j	lower case j	6/10
J	upper case J	4/10
ĵ	lower case j with circumflex accent	S 4/3 6/10
Ĵ	upper case J with circumflex accent	S 4/3 4/10
k	lower case k	6/11
K	upper case K	4/11

(Continued)

Graphic	Name or Description	Coded Representation
ķ	lower case k with cedilla	S 4/11 6/11
Ķ	upper case K with cedilla	S 4/11 4/11
ĸ	lower case k, Greenlandic	S 7/0
l	lower case l	6/12
L	upper case L	4/12
ĺ	lower case l with acute accent	S 4/2 6/12
Ĺ	upper case L with acute accent	S 4/2 4/12
ľ or ľ	lower case l with caron	S 4/15 6/12
Ľ̌ or Ľ	upper case L with caron	S 4/15 4/12
ļ	lower case l with cedilla	S 4/11 6/12
Ļ	upper case L with cedilla	S 4/11 4/12
ł	lower case l with stroke	S 7/8
Ł	upper case L with stroke	S 6/8
ŀ	lower case l with middle dot	S 7/7
Ŀ	upper case L with middle dot	S 6/7
m	lower case m	6/13
M	upper case M	4/13
n	lower case n	6/14
N	upper case N	4/14
ń	lower case n with acute accent	S 4/2 6/14
Ń	upper case N with acute accent	S 4/2 4/14

(Continued)

Graphic	Name or Description	Coded Representation
ñ	lower case n with tilde	S 4/4 6/14
Ñ	upper case N with tilde	S 4/4 4/14
ň	lower case n with caron	S 4/15 6/14
Ň	upper case N with caron	S 4/15 4/14
ņ	lower case n with cedilla	S 4/11 6/14
Ņ	upper case N with cedilla	S 4/11 4/14
ŋ	lower case eng, Lapp	S 7/14
Ŋ	upper case eng, Lapp	S 6/14
ŉ	lower case n with apostrophe	S 6/15
o	lower case o	6/15
O	upper case O	4/15
ó	lower case o with acute accent	S 4/2 6/15
Ó	upper case O with acute accent	S 4/2 4/15
ò	lower case o with grave accent	S 4/1 6/15
Ò	upper case O with grave accent	S 4/1 4/15
ô	lower case o with circumflex accent	S 4/3 6/15
Ô	upper case O with circumflex accent	S 4/3 4/15
ö	lower case o with diaeresis or umlaut mark	S 4/8 6/15
Ö	upper case O with diaeresis or umlaut mark	S 4/8 4/15
õ	lower case o with tilde	S 4/4 6/15
Õ	upper case O with tilde	S 4/4 4/15

(Continued)

Graphic	Name or Description	Coded Representation
ő	lower case o with double acute accent	S 4/13 6/15
Ő	upper case O with double acute accent	S 4/13 4/15
ō	lower case o with macron	S 4/5 6/15
Ō	upper case O with macron	S 4/5 4/15
œ	lower case œ ligature	S 7/10
Œ	upper case Œ ligature	S 6/10
ø	lower case o with slash	S 7/9
Ø	upper case O with slash	S 6/9
p	lower case p	7/0
P	upper case P	5/0
q	lower case q	7/1
Q	upper case Q	5/1
r	lower case r	7/2
R	upper case R	5/2
ŕ	lower case r with acute accent	S 4/2 7/2
Ŕ	upper case R with acute accent	S 4/2 5/2
ř	lower case r with caron	S 4/15 7/2
Ř	upper case R with caron	S 4/15 5/2
ŗ	lower case r with cedilla	S 4/11 7/2
Ŗ	upper case R with cedilla	S 4/11 5/2
s	lower case s	7/3
S	upper case S	5/3

(Continued)

Graphic	Name or Description	Coded Representation
ś	lower case s with acute accent	S 4/2 7/3
Ś	upper case S with acute accent	S 4/2 5/3
ŝ	lower case s with circumflex accent	S 4/3 7/3
Ŝ	upper case S with circumflex accent	S 4/3 5/3
š	lower case s with caron	S 4/15 7/3
Š	upper case S with caron	S 4/15 5/3
ş	lower case s with cedilla	S 4/11 7/3
Ş	upper case S with cedilla	S 4/11 5/3
ß	lower case sharp s, German	S 7/11
t	lower case t	7/4
T	upper case T	5/4
ť or t'	lower case t with caron	S 4/15 7/4
Ť	upper case T with caron	S 4/15 5/4
ţ	lower case t with cedilla	S 4/11 7/4
Ţ	upper case T with cedilla	S 4/11 5/4
ŧ	lower case t with stroke	S 7/13
Ŧ	upper case T with stroke	S 6/13
þ	lower case thorn, Icelandic	S 7/12
Þ	upper case thorn, Icelandic	S 6/12
u	lower case u	7/5
U	upper case U	5/5

(Continued)

Graphic	Name or Description	Coded Representation
ú	lower case u with acute accent	S 4/2 7/5
Ú	upper case U with acute accent	S 4/2 5/5
ù	lower case u with grave accent	S 4/1 7/5
Ù	upper case U with grave accent	S 4/1 5/5
û	lower case u with circumflex accent	S 4/3 7/5
Û	upper case U with circumflex accent	S 4/3 5/5
ü	lower case u with diaeresis or umlaut mark	S 4/8 7/5
Ü	upper case U with diaeresis or umlaut mark	S 4/8 5/5
ũ	lower case u with tilde	S 4/4 7/5
Ũ	upper case U with tilde	S 4/4 5/5
ŭ	lower case u with breve	S 4/6 7/5
Ŭ	upper case U with breve	S 4/6 5/5
ű	lower case u with double acute accent	S 4/13 7/5
Ű	upper case U with double acute accent	S 4/13 5/5
ů	lower case u with ring	S 4/10 7/5
Ů	upper case U with ring	S 4/10 5/5
ū	lower case u with macron	S 4/5 7/5
Ū	upper case U with macron	S 4/5 5/5
ų	lower case u with ogonek	S 4/14 7/5
Ų	upper case U with ogonek	S 4/14 5/5
v	lower case v	7/6
V	upper case V	5/6

(Continued)

Graphic	Name or Description	Coded Representation
w	lower case w	7/7
W	upper case W	5/7
ŵ	lower case w with circumflex accent	S 4/3 7/7
Ŵ	upper case W with circumflex accent	S 4/3 5/7
x	lower case x	7/8
X	upper case X	5/8
y	lower case y	7/9
Y	upper case Y	5/9
ý	lower case y with acute accent	S 4/2 7/9
Ý	upper case Y with acute accent	S 4/2 5/9
ŷ	lower case y with circumflex accent	S 4/3 7/9
Ŷ	upper case Y with circumflex accent	S 4/3 5/9
ÿ	lower case y with diaeresis or umlaut mark	S 4/8 7/9
Ÿ	upper case Y with diaeresis or umlaut mark	S 4/8 5/9
z	lower case z	7/10
Z	upper case Z	5/10
ź	lower case z with acute accent	S 4/2 7/10
Ź	upper case Z with acute accent	S 4/2 5/10
ž	lower case z with caron	S 4/15 7/10
Ž	upper case Z with caron	S 4/15 5/10
ż	lower case z with dot	S 4/7 7/10
Ż	upper case Z with dot	S 4/7 5/10

Diacritical Marks

Graphic	Name or Description	Coded Representation
´	acute accent	S 4/2 2/0
`	grave accent	S 4/1 2/0
^	circumflex	S 4/3 2/0
~	tilde	S 4/4 2/0
¨	diaeresis or umlaut mark	S 4/8 2/0
¸	cedilla	S 4/11 2/0
˛	ogonek	S 4/14 2/0
˘	breve	S 4/6 2/0
ˇ	caron	S 4/15 2/0
˝	double acute accent	S 4/13 2/0
˙	dot	S 4/7 2/0
¯	macron	S 4/5 2/0
˚	ring	S 4/10 2/0

Note: *The grave accent, circumflex, and tilde marks are also coded as 6/0, 5/14, and 7/14 respectively.*

1. American National Standards Institute (1983). *Videotex/Teletext Presentation Level Protocol Syntax.* Washington, D.C., X3.110-1983.